Building Croatian Medical Dictionary from Medical Corpus

Kocijan, Kristina; Kurolt, Silvia; Mijić, Linda

Source / Izvornik: **Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 2020, 46, 765 - 782**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.31724/rihjj.46.2.17

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:854666

UDK 811.163.42’373.46:61
811.163.42’322
Prethodno priopćenje
Rukopis primljen 15. IX. 2019.
Prihvaćen za tisak 23. I. 2020.
doi.org/10.31724/rihjj.46.2.17

Kristina Kocijan
Faculty of Humanities and Social Sciences, University of Zagreb
Ulica Ivana Lučića 3, HR-10000 Zagreb
krkocijan@ffzg.hr

Silvia Kurolt
Faculty of Humanities and Social Sciences, University of Zagreb
Ulica Ivana Lučića 3, HR-10000 Zagreb
skurolt@ffzg.hr

Linda Mijić
Department of Classical Philology, University of Zadar
Ulica Mihovila Pavlinovića 1, HR-23000 Zadar
lmijic@unizd.hr

BUILDING CROATIAN MEDICAL DICTIONARY FROM MEDICAL CORPUS

The overall objective of this project is to define linguistic models at the lexical and syntactic levels that appear in the health domain, depending on the type of corpus. In the first phase of the project, the texts forming the medical corpus A – MedCorA (2,232 pharmaceutical instructions for medicaments available in Croatia) were prepared. The terminology found in this corpus was analyzed and the semantic subdomains (anatomy, condition, microorganism, chemistry, etc.) within the medical domain were defined and added to the dictionary entries. These dictionary resources were used as the foundation for the second phase in which NooJ morphological grammars were built allowing annotation of medical terminology in the corpus. Said grammars were built to allow for recognizing Latinisms, as well as Latin expressions written with Croatian case endings, not only Croatian words. Prepared resources are made available to a broader scientific community via Sketch Engine for further research in the field of medicine enabling additional research and development of algorithms for, among others, medical documents classification, medical texts’ information retrieval or machine translation of medical documentation, taking into account quality and reliability as well as terminology variability.

1. Introduction

So far, the 21st century has been characterized by immense collections of unstructured data that are proving to be a real challenge from the NLP perspective. However, if we were to find a way to automatically process and understand this data in the medical domain, it could help improve analytical abilities in medical care, both at the individual and macro levels. The importance of NLP in the medical domain was acknowledged by the Institute of Medicine in 2003 (Friedman 2005) which, even at those early years of merging the NLP tools with medical data, managed to anticipate its strength in processing large volumes of text and bringing meaning to it in a timely manner. Furthermore, access to large collections of digital health records, as well as the creation of information retrieval tools, would greatly assist in conducting expensive individual clinical trials as it would offer a greater variety of samples, as well as faster and more relevant information available to those who need it. Benefits of this more accessible and more easily shared data would affect many different aspects of medical care: from making decisions on patient's treatment, adjusting a drug to a particular population, to business decisions of health institutions (Friedman 2005; Demner-Fushman et al. 2009).

Two of the probably most troublesome drawbacks of such digital records at this time fall in the domain of legal debates from the perspective of privacy regulations when data is included (Hoffman and Podgurski 2012; Birnhack 2013) or excluded (Lerman 2013), and abundance of unstructured texts, considered nowadays to be one of the real challenges of big data. To tackle the later, we have embarked on a project of preparing the dictionary of medical entries. Our preliminary corpus for extraction of such terms consists of medical entities found in pharmaceutical instructions on medicaments.

In the following sections, we will give a short background of the topic before we proceed with the detailed description of the corpus and electronic dictionary designed for the computer usage that will enable automatic detection of health-related concepts. Prior to the concluding remarks, we will show how semantic categories within the medical domain that we have introduced via the dictionary entries in NooJ are added to the SketchEngine environment to enhance the user search experience inside the Croatian Medical Corpus.

2. Theoretical overview

The development of different digital tools and computer linguistics has expanded the research into the medical domain to standardize its terminology and nomenclature. We can find in literature a variety of tools, as well as approaches for identification and classification of medical entities, mostly for English. For example, Sager et al. (1994) describe The Linguistic String Project (LSP) as one of the earliest versions of a medical NLP system that later matured into Medical Language Processor (MLP), developed for processing the narrative portions of patient records and mapping the information elements into a database representation with the overall success rate of about 85%. Grover et al. (2002) highlight the central role XML mark-up and XML NLP tools have played in the analysis and description of the resultant annotated corpus of MEDLINE abstracts. McInnes, Pedersen, and Pakhomov (2007) use a model-fitting method based on the Log Likelihood Ratio to classify three-word medical terms as right or left-branching to resolve a problem of structural ambiguity. In the same year, Sahay et al. (2007) describe a medical knowledge annotation and acquisition system called SENTIENT-MD ("Semantic Annotation and Inference for Medical Knowledge Discovery"). Gobbel (2014) proposes a tool RapTAT designed to assist annotation with the purpose of reducing the time and/or effort required to annotate a document. Savkov et al. (2016) describe a de-identified corpus of free-text notes typed by physicians during patient consultations, a shallow syntactic and named entity annotation scheme for this kind of text, and an approach to training domain specialists with no linguistic background to annotate the text. They present a statistical chunking system for such clinical texts. Christen, Groß, and Rahm (2016) propose two approaches for semantically annotating medical documents: linguistic-based (annotation mapping between a form and ontology by comparing each question of the form with the synonyms or labels of each concept from ontology) and reuse-based annotation approach that utilizes previous annotations to annotate similar medical documents.

Friedman (2005) provides an extensive list of clinical applications that are using NLP technology since early 1990. The list includes an overview of the clinical domains (*progress notes*, *surgical notes*, *pathology*, *radiology*, *discharge summary*, *admission diagnoses*, *biomedical text*, *emergency medicine*, *histopa-*

thology, *clinical reports*) and the types of application used (*quality assessment*, *encoding and improving browsing*, *interpreting findings*, *SNOMED encoding*, *ICD-10 encoding*, *assessing severity of pneumonia*, *data mining*, *extracting findings cancer-related*, *knowledge acquisition*, etc.).

A seven-step spiral approach is used for the detection of medical named entities in Arabic texts (Boujelben et al. 2011). The authors have used NooJ linguistic environment for all seven stages of the project, including corpus preparation, dictionary (with 3,365 simple medical nouns and 2,382 compound medical nouns), and morphological and syntactic grammar designs. Semantic categories they were interested in detecting align to a degree with ours like *diseases*, *symptoms*, *medical exams*, *drugs*, and in addition include *doctors*, *health institutions*, and *medical plants.* Similarly, we find NooJ in the project proposing the lexicon-grammar method for the automatic extraction and identification of medical entities in Italian texts (di Buono et al. 2015). The semantic categories of nouns used in this dictionary include *disease*, *drug*, *body parts*, *internal body parts*, *medical branch*, *symptom*, *test*, and *treatment.* Their dictionary was based on the Italian clinical texts comprising of 989 medical records with 41,409 different tokens. Existing databases of medical entities can be the basis for building different dictionary types depending on the specificities of an intended project.

3. Corpus

The medical corpus MedCor is made up of 6,500 pharmaceutical instructions on medicaments written for the medical personnel and available via the *Mediately* website. The first subsection of the medical corpus is named MedCor**A** and it consists of 2,232 texts (i.e. files) with 71,911,667 tokens in total (i.e. 531,672 different tokens). Of that, 30,899,536 are word forms.

Each file is an instruction for one specific medicament available in Croatia. We have used the ATC codes (Anatomical Therapeutic Chemical Classification System) for the naming system of each text file. The ATC seven-character alphanumeric code consists of 1 letter followed by 2 digits, then 2 letters, and finally another 2 digits. This coding system is used worldwide to classify different medicaments by their active ingredient, system on which they act, and chemical properties they have (WHOCC, nd).

Some medicaments come in different packaging sizes and versions but still have the same ATC code. To differentiate between such files, we have added the dosage (usually in micrograms) after the ATC identification, dividing them with an underscore "_". There are also cases when the same medicine name has different manufacturers. Just like with different dosages, the manufacturer's name is added after the ATC code followed by an underscore (e.g. *N05AX12_Cipla*, *N05AX12_Pliva*, *N05AX12_Stada*). In those cases where medicaments have the same ATC identification but go by a different name, the name is added at the end (written in all lower case letters) (e.g. *J01DC02_axetine*, *J01DC02_novocef*).

Each text file has a similar structure and includes information such as the name of the medicament, its qualitative and quantitative ingredients, pharmaceutical form, clinical data (purposes, uses, dosages, contraindications, special warnings, interactions with other medicaments, side effects), pharmacological attributes, pharmaceutical information and some marketing information. Among the nouns found in the corpus, little over 62% fall in the medical domain and they mostly follow the same pattern of distribution as found in the dictionary (for more details see Figure 1 in Section 4). The only sub-categories that show different patterns (i.e. they are found more frequently in the corpus) are the *measure* and *patient* sub-categories, which is justified by the type and nature of the corpus used.

4. Dictionary

Croatian NooJ dictionary holds 21,071 noun entries with all of their appropriate number and case combinations (Vučković et al. 2010). However, the dictionary itself did not prove adequate in the analysis of the **MedCor** corpus. This was to be expected since the dictionary covers mostly nouns found in the general corpus. The first linguistic analysis recorded a little over 88 thousand unknown words. This list was populated with different cases of same words (e.g. *abdomen*, *abdomena*, *abdomenom*, *abdomenu*), but also with some misspelled forms (e.g. *abodemnu*, *abdomenup*).

After removing duplicates and incorrect forms, we were able to extract 2,373 nouns, belonging mainly to the medical domain that were originally not includ-

ed in our dictionary of Croatian nouns. The remaining unknown words belong to other word categories from the medical domain, namely adjectives and verbs, but also some adverbs. However, at this time, we chose to focus only on nouns as our target words, leaving the other part-of-speech categories for the later stages of the project.

Although the decision to describe only nouns seemed to be quite straightforward at first, deeper analysis proved this assumption quite wrong. Thus we made another subdivision of target words that included single word units, abbreviations[1] (including ones with case endings after the "-" sign, e.g. "*MRA*", but also its instrumental case[2] form "*MRA-om*") and finally, single word units in Latin (Kocijan et al. 2019). Multiword units (MWU), like 'oral contraceptives' or 'health worker' were not processed at this time. MWU such as '*protonska pumpa*' or '*gastrointestinalnom sustavu*', where the noun (*pumpa*, *sustav*) is usually not associated with medical terminology, are not tagged at this time but will be dealt with after all the adjectives have been annotated in the NooJ dictionary. However, if a multiword includes a noun (or any number of nouns) from the medical domain (e.g. 'contraception' in 'oral contraception' or 'muscle' and 'spasm' in 'muscle spasm'), each noun is annotated as if it was a single noun.

Since we have previously started working on Latin phrases found in the medical texts (Kocijan et al. 2019), we have included them in this part of the project as well. Latin nouns, in addition to all the semantic categories as prepared for Croatian nouns, have one more (semantic) tag denoting language, as the main language of the text is Croatian. Since there are four types of notations in Croatian medical texts that refer to the same concept, three of which are based on a Latin root, each was marked differently. Thus, the pure Latin terms (i.e. *diabetes mellitus*) are marked with the tag +LAT. Latinisms, or Croatian terms with a visible Latin root (i.e. *dijabetes melitus*) are marked +LATCro. Finally, Croatinised Latin words (Latin root with Croatian case ending) (i.e. *diabetes mellitus**om**)* are marked +CROLat. Croatian translations of these terms (i.e. *šećerna bolest*) carry

[1] Abbreviations were included at this time since they are reported to be quite frequently used in clinical notes (Friedman 2005).

[2] In declension of Croatian abbreviations, a hyphen is inserted between abbreviation and case suffix (e. g. *JIL-u*, *EEG-u*, *HIV-a*, *NMS-a*). Most abbreviations are declined as male nouns (gen. *-a*, dat. *-u*, instr. *-om*) and abbreviations ending in *-a* are female (gen. *-e*, dat. *-i*, acc. *-u*, instr. *-om*). Abbreviations are written without a hyphen if they are lexicalized (e.g. *sidu* < *SIDA*) (Gjuran-Coha and Bosnar-Valković 2008).

no language marker in this corpus since it is originally a Croatian text and this information would thus be redundant.

After the work on the dictionary was completed, i.e. unknown nouns were added and described with their grammatical tags (part of speech, type, gender, inflectional paradigm) and semantic tags (medical domain and subdomain(s)), we were left with the dictionary of 27,452 nouns, 6,925 of which belong to the medical domain. It is important to note that our original dictionary did in fact hold some nouns that were also from the medical domain, but those were mostly ones likely to be found in everyday usage i.e. *arm*, *head*, *heart*, etc. We were able to manually locate those nouns and enhance their dictionary description with the data on domain and subdomain.

4.1. Semantic categories

The analysis of the MedCor corpus has revealed different semantic categories of words found in this type of text. We have opted to use a classification system of 12 categories[3] (Table 1) that may turn out to be useful for future text mining projects. At this time, we wanted to keep a more general subgrouping, while the greater degree of granularity can be added at later stages depending on the needs of future medical-based projects.

Table 1: List of medical subdomain categories, with description, examples and number of dictionary entries

Category	Description	Examples	# of dictionary occurrences
+ANAT ***[anatomy]***	any part of anatomy	*hand*, *kidney*, *veins*	621
+COND ***[condition]***	any type of disease, disorder or condition, either physiological or psychological	*headache*, *flue*, *neuralgia*	2,069

[3] Compare to 17,000 codes in The International Classification of Disease, 9th revision, Clinical Modification (ICD-9-CM) and even more than 155,000 codes in its 10th revision ICD-10-CM; or to other codes used in different clinical vocabularies like The Systematized Nomenclature of Medicine (SNOMED), Logical Observation Identifiers Names and Codes (LOINC), Unified Medical Language System (UMLS), to name the few (Steele 2017). Notice that each is using a different coding system.

Category	Description	Examples	# of dictionary occurrences
+DISCIPL *[discipline]*	any branch within medical domain	*haemathology*, *histology*, *embryology*	90
+DRUG *[drug]*	any brand name of a medicament	*Synopen*, *Viranti*	55
+KEM *[chemistry]*	any chemical, solid or liquid medicament, chemical and generic names of medicaments	*sodium*, *calcium*, *apaurin*, *aspirin*	3,167
+MEASURE	any metric, imperial and other units of measure	*mmol*, *min*, *INR*	44
+MICROORG *[microorganism]*	any micro-organism including viruses, bacteria, fungi, spores	*lactobacillus*, *trichophyton*	70
+PATIENT	any type of patient	*diabetic*, *asthmatic*	78
+PLACE	any place within health institutions	*laboratory*, *pediatrics*, *radiology*	35
+PROC *[procedure]*	any medical procedure performed within health domain	*operation*, *biopsy*, *hemodialysis*, *electrocardiogram*	610
+PROF *[profession]*	health professions	*pediatrician*, *nurse*, *doctor*, *anesthesiologist*	113
+TOOL	any tool used for performing or assisting with medical procedures	*microscope*, *stetoscope*, *sphygmomanometer*, *otoscope*, *thermometer*	110

These semantic codes have been added manually to dictionary entries denoting a subdomain within the medical domain (e.g. *Domena=MED+DomenaType =KEM*). There are some cases of nouns being able to take on more than just one semantic tag belonging to the medical domain. In such cases, a noun is marked with both *DomainType* tags. Such examples are *ultrazvuk* (*ultrasound*) used both as a procedure and as a tool, *agonist* (*agonist*) that is a part of human anatomy, but also a chemical that binds itself to a receptor and activates it, and *bradavica* that is both a part of human anatomy (*nipple*) and a medical condition (*wart*).

At the moment, medical nouns make up a little over 25% of all nouns in the dictionary. The distribution of medical subdomains inside the dictionary is visual-

ized in Figure 1 (solid line). The largest number of medical nouns falls into the *chemistry* subdomain, followed by the *condition* subdomain. This is not surprising due to the type of corpus we are working on (i.e. descriptions of medicaments). It is also to be expected that this distribution would be different if our corpus included patients' health records or medical publications.

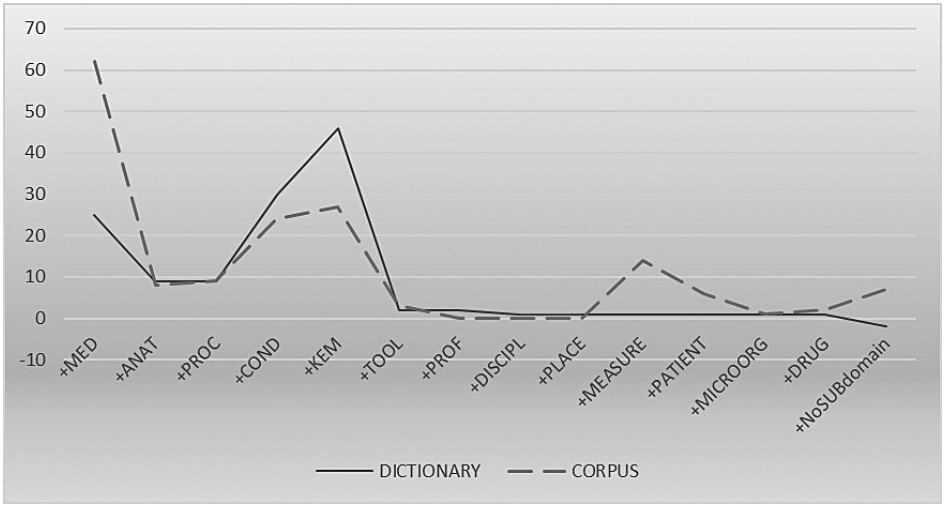

Figure 1: Distribution of medical domain and subdomain words in the dictionary (*solid line*) vs. the MedCor corpus (*dashed line*)

4.2. Ambiguities

Problems we encountered during the semantic annotation process fall into two main categories of ambiguity: within the medical domain (inner ambiguity), and outside of it (outer ambiguity). This is a common problem for many languages, thus we believe it is important to explain how we dealt with it in the Project.

If we take, for example, the noun *koncentracija (concentration)* found in the following sentences in the corpus:

a) *povećava se* ***koncentracija*** *fenitoina* [*phenytoin* ***concentration*** *level is increased*]

b) *povećava pažnju i* ***koncentraciju*** [*increases focus and* ***concentration***]

we find two words, belonging to the same part of speech – nouns, but with different meanings. In the former example, the word *concentration* carries the meaning 'the amount of a component in a given area or volume' (Merriam-Webster dictionary entry) while in the latter it means 'the ability to think carefully about something you are doing and nothing else' (Cambridge Dictionary entry). Concerning our categories, the first example would fall within the *chemistry* subdomain, while the second one talks about a *condition* as defined in our list of semantic categories (Table 1). Similarly, the word *bradavica* (*nipple/wart*) is found in two contexts in the corpus:

c) *gel se ne smije nanositi na **bradavice** žena tijekom dojenja kako bi se lijek na taj način primijeni djetetu [women who breastfeed should not try and administer the drug to the child by applying the gel to the **nipples** during the breastfeeding period]*

d) *madeži, **bradavice** i sve ostale promjene i upale na koži… [moles, **warts** and all other skin changes and inflammations…].*

In example **c**, the word *bradavica* refers to a part of the body and thus carries the semantic tag for *anatomy*, while in example **d** it should be marked as a *condition*.

To be able to find both meanings, we decided to introduce both categories to the same dictionary entry, or in our examples:

koncentracija,N+c+f+Domena=MED+DomenaType=KEM
+DomenaType=COND+FLX=MEDO

bradavica,N+c+f+Domena=MED+DomenaType=ANAT
+DomenaType=COND+FLX=BRNJICA.

These ambiguities will be sorted out at the syntactic level of analysis, which remains to be done at later stages of the project. Before the final description, examples, i.e. categories, were checked against several on-line dictionaries (MSD priručnik, Hrvatska enciklopedija, Hrvatski jezični portal, Struna – Hrvatsko strukovno nazivlje) to make sure we do not introduce erroneous tags.

Medical domain nouns with multiple different subdomains aside, we also encountered ambiguities between words belonging to different part-of-speech categories, as can be seen in the following examples:

e) *Pumpajući krv kroz arterije stvara niz tlačnih valova koje nazivamo puls ili **bilo**. [By pumping blood through arteries, it creates a series of pressure waves called pulse or **heartbeat**.]*

f) *...kako bi na vrijeme sve **bilo** dovršeno i **bilo** spremno za izlaganje. [...in order for everything to **be** finalized and ready to **be** presented on time.]*

The word *bilo* in example **e** clearly refers to a noun belonging to the medical domain. However, the occurrence of the same word in the example sentence **f** denotes an auxiliary verb 'to be'.

This situation had to be sorted out by introducing two different dictionary entries, one for the noun and the other for the verb, therefore relying on the lexical analysis to disambiguate between the two tags.

5. Flirting with the SketchEngine

SketchEngine has become a well-known platform for exploring different corpora, thanks in part to its language variety and availability to students and researchers worldwide. Since these resources are currently available for free, we decided to use its potential and make our resources available to the rest of the research community. Due to the specialized nature of the texts we had, we named the corpus *CMC – Croatian Medical Corpus*.

In the following sections, we will describe the steps for building a specialized corpus with additional semantic tags. It was a three-step process that can in short be described as: Step 1: prepare the corpus in SketchEngine; Step 2: Add semantic tags in NooJ; Step 3: Return tagged corpus to SketchEngine.

5.1. Building the CMC corpus

There are several ways to build a corpus in SketchEngine, one of which being via upload of PDF files. Since our texts were originally obtained as PDF files, this was the option that worked best for us. Upon the successful file upload, the uploaded text files need to be compiled. During this phase, SketchEngine annotates the text files with a predefined tag set for the language in use, which in

our case meant the Croatian tag set (Erjavec and Ljubešić 2016). The result is a vertical file with 4 types of data (or 4 different columns):

1) the word, as found in the text,
2) its grammatical description depending on the part-of-speech category (part-of-speech category itself, gender, number, etc.),
3) its base form followed by a dash and part-of-speech category, and
4) its base form written in all lowercase letters.

However, if the corpus consists of multiple files, as is the case with *CMC* (each file is a description of one medicament), they get compiled and verticalized into one single file. This was not desirable, as we wanted to keep the original files and the file naming system. Thus, we opted to upload one file at a time, compile it on its own and then download it as a vertical file. We repeated these steps for each of the files in the corpus.

5.2. Adding the semantic tags

Verticalized files can be downloaded and opened in Microsoft Excel where each vertical row equals one row of an Excel spreadsheet. This made it fairly easy to simply copy the first row (populated with the original unmarked text) and import it into the NooJ environment where we prepared our original dictionaries with semantic tags for the medical domain and its subdomains. We then performed an automatic linguistic analysis of this imported text and exported it back into Excel afterwards. The exported text is again a vertical list, but this time it contains additional semantic annotations. Since the original and annotated texts were aligned, it was quite straightforward to add the annotations produced in NooJ to the original vertical file before returning it to SketchEngine. But before doing so, we performed a manual double-check of the annotations to make sure no errors were present in the files. During this process, we noticed multiple erroneous tags that were not added manually but were instead introduced by SketchEngine in the process of automatic annotation. These newly introduced errors can be divided into two categories: either a different part-of-speech category was incorrectly tagged as a noun (Table 2), or a noun was incorrectly tagged as another part-of-speech category (Table 3).

For example, the word *adenomatoidna* is an adjective but is marked by SketchEngine as a noun (Table 2). Although it belongs to the medical domain, we chose

not to mark it for the semantic medical domain at this time, as it is not a noun regardless of the tag given to it by SketchEngine. This is true for all other incorrectly marked (non)nouns.

Table 2: The adjective 'adenomatoidna' marked incorrectly as a noun

Word	Tags	Word-POS	Word in lowercase letter	Domain	Subdomain
adenomatoidna	Ncnsg	adenomatoidno-n	Adenomatoidno		
Hiperplazija	Ncfpg	hiperplazija-n	Hiperplazija	MED	COND

In another example, noun *omeprazola* is also annotated both as an adverb and as a verb – out of the 125 occurrences of the noun in the same text, 2 are marked as a verb, 5 as an adverb, 4 as an adjective and remaining 114 as a noun. It should also be noted that the form of the noun itself is the same in each case of mis-tagging, and the erroneous tags seem to be dependent on the context of the noun, rather than the noun itself. Still, we decided to add the semantic tags to all of these occurrences regardless of whether they are carrying the wrong part-of-speech tag, as they are still nouns regardless of the tag. However, this will result in other part-of-speech categories to have the semantic annotations as well, which may give the SketchEngine user incorrect information on the data distribution.

Table 3: Same noun with different part-of-speech tags (marked in column 2 – 'Tags')

Word	Tags	Word-POS	Word in lowercase letter	Domain	Subdomain
Example 1					
Doziranje	Ncnsv	doziranje-n	doziranje	MED	PROC
Omeprazola	Rgp	omeprazola-r	omeprazole	MED	KEM
Example 2					
Klirens	Ncmsn	klirens-n	klirens	MED	PROC
Omeprazola	Vmp-sf	omeprazoti-v	omeprazoti	MED	KEM
Example 3					
metaboliziranje	Ncnsa	metaboliziranje-n	metaboliziranje	MED	PROC
omeprazola	Ncmsg	omeprazol-n	omeprazole	MED	KEM

5.3. Returning data to SketchEngine

The Excel spreadsheet with both the original and newly added tags was again saved as a vertical file (Figure 2) that was then re-uploaded to SketchEngine. To read and recognize the added tags from this new file, a small corpus template needed to be designed where the new tags were defined.

```
2 SAŽETAK Npmsn   sažetak-n   sažetak
3 OPISA   Npmsn   opis-n  opis
4 SVOJSTAVA   Aspfsny svojstvo-a  svojstvo
5 LIJEKA  Npfsn   lijek-n lijek   MED KEM
6 1   Mdc [number]-m  [number]
```

Figure 2: An excerpt of a vertical file after adding the semantic annotations for medical domain and subdomains

This was done via the original SketchEngine interface since this feature is still not supported in the new version. Nevertheless, the *Croatian Medical Corpus* is fully available via the new interface with all the options available both for finding words (e.g. via CQL – Corpus Query Language) and for showing the results in KWIC that may include any selection of attributes of found concordances (Figure 3).

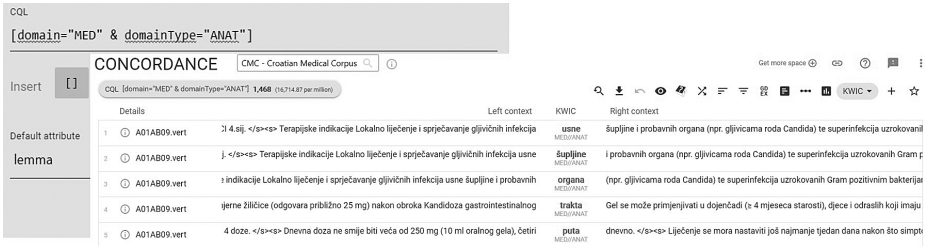

Figure 3: Results of a CQL query done in SketchEngine – searching for all words that have a medical domain and anatomy subdomain

The selection of results presented in Figure 3 was obtained after the CQL query [domain=“MED” & domainType=“ANAT”] was executed. The concordance screen shows the name of the vertical file where the data is found, left and right context of the token, as well as the 2 attributes that denote the searched domain (MED) and subdomain (ANAT).

6. Conclusion and future work

Any NLP project conducted within the health domain is a valuable endeavor, albeit not an easy one. This difficulty is amplified with the fragility of any human life that such research may have an impact on. Thus, we have approached our data with the great responsibility for making the correct classification of subdomains and rechecking the algorithm results manually, before the annotated text is made available via SketchEngine. The process we have employed for detecting semantic categories within the medical domain, including the ambiguity resolution, is described in detail. The same was done for the steps taken to prepare the annotated text for SketchEngine search and manipulation.

Still, regardless of the amount of effort already devoted to the project, the process is only in its infancy and requires further work. After completing the semantic annotation of nouns, we will proceed with other parts of speech that may carry medical-related meaning, before we can start detecting and annotating multi-word expressions. Due to the specificity of files used by SketchEngine, the medical-MWE detection will need a slightly different approach that will require additional algorithms to be designed within the NooJ environment to allow faster automatic annotations of such expressions.

References

Birnhack, Michael. 2013. S-M-L-XL Data: Big Data as a New Informational Privacy Paradigm. *Big Data and Privacy: Making Ends Meet 7-10 (Future of Privacy Forum & Center for Internet & Society)*. Stanford Law School. Stanford. dx.doi.org/10.2139/ssrn.2310700.

Boujelben, Ines; Mesfar, Slim; Ben Hamadou, Abdelmajid. 2011. Methodological approach of terminological extraction applied to biomedical domain. *Proceedings of the 4th International Conference on Information Systems and Economic Intelligence SIIE2011.* Marrakech.

Christen, Victor; Gross, Anika; Rahm, Erhard. 2016. Approaches for Annotating Medical Documents. *LWDA*. 227–232.

Demner-Fushman, Dina; Chapman, Wendy W.; McDonald, Clement J. 2009. What can natural language processing do for clinical decision support?. *Journal of Biomedical Informatics* 42. 760–772.

di Buono, Maria Pia; Maisto, Alessandro; Pelosi, Serena. 2015. From Linguistic Resources to Medical Entity Recognition: A Supervised Morpho-syntactic Approach. *Proceedings of the ALLDATA2015: The First International Conference on Big Data, Small Data, Linked Data and Open Data*. Eds. Grzymala-Busse, Jerzy W.; Schwab, Ingo; di Buono, Maria Pia. International Academy, Research, and Industry Association. Barcelona. 81–86.

Friedman, Carol. 2005. Semantic text parsing for patient records. *Knowledge management and data mining in biomedicine*. Eds. Chun, Hsincun et al. Springer. New York. 423–448.

Gjuran-Coha, Anamarija; Bosnar-Valković, Brigita. 2008. Uporaba kratica u jeziku medicinske struke. *Filologija* 50. 1–12.

Gobbel, Glenn; Garvin, Jennifer H.; Reeves, Ruth Madeleine; Cronin, Robert M.; Heavirland, Julia; Williams, Jenifer; Weaver, Allison; Jayaramaraja, Shrimalini; Giuse, Dario; Speroff, Theodore; Brown, Steven H.; Xu, Hua; Matheny, Michael E. 2014. Assisted annotation of medical free text using RapTAT. *Journal of the American Medical Informatics Association* 21/5. 833–841. doi.org/10.1136/amiajnl-2013-002255.

Grover, Claire; Klein, Ewan; Lapata, Mirella; Lascarides, Alex. 2002. XML-based NLP tools for analysing and annotating medical language. *Proceedings of the 2nd Workshop on NLP and XML (NLPXML-2002)*. Eds. Wilcock, Graham; Ide, Nancy; Romary, Laurent. Taipei. 29–36.

Hoffman, Sharona; Podgurski, Andy. 2012. Big Bad Data: Law, Public Health, and Biomedical Databases. *Journal of Law, Medicine and Ethics*: *Case Legal Studies Research Paper* 34. 56–60. doi.org/10.1111/jlme.12040.

Kocijan, Kristina; di Buono, Maria Pia; Mijić, Linda. 2019. Detecting Latin-based Medical Terminology in Croatian Texts. *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications*. Eds. Mirto, Ignazio Mauro; Monteleone, Mario; Silberztein, Max. Springer International Publishing. Palermo. 38–49.

Lerman, Jonas. 2013. Big Data and Its Exclusions. *Stanford Law Review Online* 66. dx.doi.org/10.2139/ssrn.2293765.

McInnes, Bridget; Pedersen, Ted; Pakhomov, Serguei V. 2007. Determining the syntactic structure of medical terms in clinical notes. *Biological, translational, and clinical language processing*. Association for Computational Linguistics. Prague. 9–16. www.aclweb.org/anthology/W07-1002 (accessed 21 June 2020).

Sager, Naomi; Lyman, Margaret; Bucknall, Christine; Nhan, Ngo; Tick, Leo J. 1994. Natural Language Processing and Representation of Clinical Data. *Journal of the American Medical lnformatics Association* 1/2. 142–160.

Sahay, Saurav; Agichtein, Eugene; Li, Baoli; Garcia, Ernest V.; Ram, Ashwin. 2007. Semantic annotation and inference for medical knowledge discovery. *Proceedings of the NSF Symposium on Next Generation of Data Mining*.

Savkov, Aleksandar; Carroll, John; Koeling, Rob; Cassell, Jackie. 2016. Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus. *Language Resources and Evaluation* 50/3. 523–548.

Steele, Marsha C. 2017. *Classification systems, clinical vocabularies, and terminology*. Nurse Key. https://nursekey.com/6-classification-systems-clinical-vocabularies-and-terminology/ (accessed 21 June 2020).

Vučković, Kristina; Tadić, Marko; Bekavac, Božo. 2010. Croatian Language Resources for NooJ. *CIT. Journal of computing and information technology* 18/4. 295–301. doi.org/10.2498/cit.1001914.

WHOCC = WHO Collaborating Centre for Drug Statistics Methodology. 2018. www.whocc.no/atc_ddd_methodology/history (accessed 21 June 2020).

E-dictionaries

Cambidge Dictionary. Cambidge Dictionary Press, 2020. https://dictionary.cambridge.org (accessed 21 June 2020).

Hrvatska enciklopedija. Leksikografski zavod *Miroslav Krleža*. www.enciklopedija.hr (accessed 21 June 2020).

Hrvatski jezični portal. http://hjp.znanje.hr (accessed 21 June 2020).

Merriam-Webster.com. Merriam-Webster. https://www.merriam-webster.com (accessed 21 June 2020.)

MSD priručnik. www.msd-prirucnici.placebo.hr (accessed 21 June 2020).

Struna – Hrvatsko strukovno nazivlje. Institut za hrvatski jezik i jezikoslovlje. http://struna.ihjj.hr (accessed 21 June 2020).

Izrada hrvatskoga medicinskog rječnika iz medicinskoga korpusa

Sažetak

Osnovni je cilj ovoga projekta definiranje leksičkih i sintaktičkih jezičnih modela koji se pojavljuju u području medicine, a ovisno o vrsti korpusa. U prvoj fazi projekta prikupljeni su tekstovi koji čine medicinski korpus A – MedCorA (2232 farmaceutske upute za lijekove dostupne u Hrvatskoj). Nazivlje je iz korpusa analizirano, a potom su definirane semantičke poddomene (anatomija, stanja, mikroorganizmi, kemija itd.) unutar medicinske domene. Semantičke su oznake dodane u rječnik, gdje su poslužile kao osnova za drugu fazu projekta u kojoj su izrađene i NooJ morfološke gramatike za prepoznavanje i označavanje latinizama kao i latinskih izraza koji se koriste hrvatskim padežnim nastavcima.

Pripremljeni resursi stavljaju se na raspolaganje široj znanstvenoj zajednici putem SketchEnginea za daljnja istraživanja u području obrade jezika i medicine, omogućujući pritom nova istraživanja i razvoj algoritama za, među ostalim, klasifikaciju medicinskih dokumenata, pronalaženje podataka u medicinskim tekstovima, prevođenje medicinske dokumentacije, a uzimajući u obzir kvalitetu i pouzdanost podataka, ali i terminološku varijabilnost.

Keywords: language processing, semantic annotations, medical domain, NooJ, Croatian

Ključne riječi: obrada jezika, semantičke oznake, medicinska domena, NooJ, hrvatski jezik